下载APP

06 | jieba分词：如何基于感情色彩进行单词数量统计?

2021-02-23 尹会生

Python自动化办公实战课 进入课程 >

讲述：尹会生
时长 13:01 大小 11.93M

你好，我是尹会生。

在涉及运营、市场的工作中，我们经常需要根据产品评论的情感分析，来了解某一产品的口碑。所谓的情感分析，就是指根据用户对产品的评论，分析出用户对产品的喜好程度。

最简单的，我们会**区分产品的评价是正向还是负向**的，然后根据反馈结果改变产品的特性。稍微复杂一点的，我们会**根据情感色彩将产品的评价关键词提取出来，进行统计和分类（用于更深入的分析产品）。**

如果靠人工对产品评价进行辨析，有很大的局限性：一个是不够公平，因为每个人对词语感情色彩的理解并不是完全一致的；另一个是产品评价有很多，而且还会不定期增加，人工分析很难保证及时性。

因此，在进行词语的情感分析时，我通常都会使用 Python 的 jieba 库，来自动化实现文本情感分析功能。一般需要经过三个步骤，分别是**分词、优化分词结果和情感分析**。

那我就先带你看看**为什么要进行分词，以及如何进行分词操作。**

如何分词？

要想判断一段话表达的情感是正向还是负向，就需要根据这句话中的关键词来得到情感的倾向。例如一段话中出现了“开心”“高兴”“物超所值”等正向的词语，我们就可以认定这条产品的评价是偏正向的。相反，出现“不喜欢”“差”等词语，评价就是偏负向的。

但是，要想从一句话中将这些表达情感的词一个一个找出来，就需要依靠专业的工具把一句话根据语义划分成多个词，再把表达情感的词语提取出来，进行情感分析。

为什么要先根据语义来划分词呢？这主要是因为中文句子里的每个词中间没有用空格进行分隔，没有分隔就没法进行之后的情感分析。而对中文句子按照语义进行切割的这种操作，我们就称为“分词”。

Python 中有非常成熟的分词库，其中最流行的库是 jieba 库。在计算机中，实现语义分割的技术有两种，一种是从统计学的角度分词，另一种是从词库的角度基于 TF-IDF 算法实现分词。jieba 就是采用第二种，基于词库的角度对文章进行自动分词的。

那我就以电商网站上的一段商品评论为例，给你演示一下 jieba 库是如何实现分词的。

复制代码

```
import jieba

words1="速度快，包装好，看着特别好，喝着肯定不错！价廉物美"

words2 = jieba.cut(words1)

print("/".join(words2))
# 速度/快，/包装/好，/看着/特别/好，/喝/着/肯定/不错/！/价廉物美
```

在这段代码中，我利用 jieba 库的 cut() 函数实现了自动分词功能。我刚才讲了，jieba 分词是依靠词库实现的，词库里包含了提前准备好的词和词性。下图就是 jieba 词库的内

容：

```
一鼓 ru
一鼓作气 ru
一马当先 ru
... ...
```

这些词库中的词，jieba 是怎么识别的呢？

在你使用 pip 命令安装了 jieba 库之后，它会附带一个默认的词库。在官方文档中，将这个词库称作“字典”文件。这个文件包含了日常汉语的词语、词性。jieba 库会先基于“字典”对文章中所有可能出现的词进行匹配。匹配之后，会生成句子中的汉字所有可能形成的词。然后再将这些词构成的有向无环图（DAG），并采用动态规划算法查找最大概率路径，尽可能不会将一个词拆分成单个汉字。最后再从“字典”找出基于词频的最大切分组合，把这分词的组合从句子中找出来，形成一个一个的词。

而且，为了提高分词的准确率，jieba 对没有记录在字典的词（称作未登录词）也使用了分词的模型，它就是大名鼎鼎的基于汉字成词能力的 HMM 模型（隐马尔可夫模型）。对词库中的词和未登录词进行处理之后，jieba 就可以实现自动化分词了。

不过，分词之后，我们还需要对分词结果进行优化。因为在分词结果中存在着大量的标点符号，还有“看着”“喝着”“包装” 等和表达产品评价的情感无关的词语，为了加快计算词语的情感速度、避免无关词语影响情感倾向判断，我们就要优化分词的结果。

优化分词结果

优化分词结果主要从两个方面进行优化：一方面是移除标点符号；一方面是删除和情感无关的助词、名词等。

我先来带你学习下怎么从分词结果中移除标点符号。

移除标点符号一般有两种方法：

1. 删除停止词（Stop Words）；

2. 根据词性提取关键词。

先来看看第一种，删除停止词。

所谓的停止词，就是指**为了节省空间和提高匹配词语情感倾向的效率，在进 **** 行情感分析前自动过滤掉的某些字或词。**

停止词主要是标点符号，也可以是“啊呀呢”等语气助词。把标点符号写入停止词列表后，再使用 for 循环功能，将 jieba 分好的词和停止词列表依次匹配。如果 jieba 分好的词出现在列表中，就将这些词删掉。如果没有出现在列表中，就把这些词再组合成一个新的列表，后续就可以对新的列表进行情感分析。

删除停止词的代码如下。通过删除停止词，我们就可以得到只有汉字的分词结果。

复制代码

```
words2 = jieba.cut(words1)
words3 = list(words2)
print("/".join(words3))
# 速度/快/, /包装/好/, /看着/特别/好/, /喝/着/肯定/不错/! /价廉物美

stop_words = [", ", "! "]
words4 =[x for x in words3 if x not in stop_words]
print(words4)
# ['速度', '快', '包装', '好', '看着', '特别', '好', '喝', '着', '肯定', '不错', '价
```

另一种优化分词结果的方式叫做根据词性提取关键词。这种方式的优点在于不用事先准备停用词列表，jieba 库就能够根据每个词的词性对其进行标注。

我这里为你提供了一张 paddle（paddle 是百度开源的深度学习平台，jieba 使用了 paddle 的模型库）模式词性表作为参考，你可以根据 jieba 自动分析得到的词性结果，手动将助词、虚词（标点符号）移除。

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	t	时间
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通名词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

我把这个基于词性移除标点符号的代码也提供给你：

复制代码

```
# words5 基于词性移除标点符号
import jieba.posseg as psg
words5 = [ (w.word, w.flag) for w in psg.cut(words1) ]
# 保留形容词
saved = ['a',]
words5 =[x for x in words5 if x[1] in saved]
print(words5)
# [('快', 'a'), ('好', 'a'), ('好', 'a'), ('不错', 'a')]

```

在这段代码中，我在使用 jieba 库的 posseg 类实现分词的同时，也对词性进行了标注。为了让你看到更直接的结果，我只保留了形容词，因此，变量 saved 的列表参数就只有一个‘a’，表示保留的词类型为形容词。

如果你希望保留更多的词性，可以将词性表中代表每种词的英文缩写写入 saved 列表中，其中，我建议你在处理之后把形容词、副词、动词都保留下来，这些都有助于你进行下一步的语义情感分析。

在优化分词结果之后，我们就得到了只有形容词的处理结果。那么，接下来，我们需要基于这些形容词来获取产品评价的正向或负向结果，以及基于词语的情感色彩来统计单词的数量。

语义情感分析

对于已经分好词的语句，我们需要使用另一个库统计词的正向、负向情感倾向，这个库就是 snownlp 库。

snownlp 库既能实现分词，也能计算词出现频率，以及进行情感分析。那你可能就发出疑问了：为什么不直接使用 snownlp 进行分词，而要使用 jieba 分词呢？

原因就在于，snownlp 的算法问题，会让它对否定词划分得不够准确。例如“不喜欢”，snownlp 会把这个词划分为两个独立的词，分别是“不”和“喜欢”。那么，在计算语义情感时，就会产生较大的误差。所以我们会先采用 jieba 进行分词，分词之后再采用 snownlp 来实现语义情感分析功能。

接下来，我带你看一下如何使用 snownlp 得到完成分词之后的情感分析结果。代码如下：

复制代码

```
from snownlp import SnowNLP
words6 = [ x[0] for x in words5 ]
s1 = SnowNLP(" ".join(words3))
print(s1.sentiments)
# 0.99583439264303
```

这段代码通过 snownlp 的 Bayes（贝叶斯）模型训练方法，将模块自带的正样本和负样本读入内存之后，再使用 Bayes 模型中的 classify() 函数进行分类，这样就得到了 sentiments 属性的值，sentiments 的值表示情感倾向的方向。在 snownlp 中：

如果情感倾向是正向的，sentiments 的结果会接近 1。

如果情感倾向是负向的，结果会接近 0。

可以看到，我们在刚刚的代码中得到的情感分析的结果是 0.9958，非常接近 1，因此这条产品的评价就是正向的。

情感倾向结果趋近于 1 或者趋近于 0 都是非常理想的情况，可以直接得到感情色彩比较强烈的产品评价。但是，有时候感情色彩不太强烈，在这种情况下，我们就需要根据评价的数值范围对评论进行分组，统计每组包含多少个评价。

这个功能也可以通过 snownlp 实现，我把代码写在这里，你可以参考：

复制代码

```
positive = 0
negtive = 0
for word in words6:
    s2 = SnowNLP(word)

    if s2.sentiments > 0.7:
        positive+=1
    else:
        negtive+=1

    print(word,str(s2.sentiments))
print(f"正向评价数量:{positive}")
print(f"负向评价数量:{negtive}")
# 快 0.7164835164835165
# 好 0.6558628208940429
# 好 0.6558628208940429
# 不错 0.8612132352941176
# 价廉物美 0.7777777777777779
# 正向评价数量:3
# 负向评价数量:2
```

通过 snownlp 库配合 jieba 分词的结果，你就可以实现批量产品评论的自动语义情感分析了。同时，你还可以根据不断累积产品的评价，来持续优化你的产品。

小结

最后，我来为你总结一下对文件进行情感倾向分析的关键步骤和注意事项。实现语义情感分析功能，你必须掌握分词、优化分词结果、语义情感分析这三个步骤。

其中分词是实现中文语义分析的第一步，也是最基础的部分。分词的好坏决定了对每个词的情感进行标注的准确程度。如果默认的 jieba 分词没有正确地把词语划分，你也可以使用 jieba 自带的 suggest_freq() 函数进行词频调节。

举个小例子，“中”“将”两个字可以组成“中将”的词语，也可以拆开来用“我们中 / 将有人成功考上北大”。在不同的语言环境中，我们要通过词频调节来让它们以词的形式出现，还是以两个字的方式出现。调整的方法是：

```
jieba.suggest_freq(("中", "将"), tune = True)
```

复制代码

可以看到，利用调节词频使“中” “将”都能被分出来，而不会被错误地识别为一个词“中将”，通过这种方式，就可以提升 jieba 的识别正确率。

在优化分词结果这一步，你可以通过减少虚词和标点符号，通过停止词、词性的选择，来降低它们对情感分析结果的干扰。

最后，你还可以为 snownlp 增加新的流行词和网络用语，帮你更准确地分析用户对产品的喜好程度，从而提高产品定位的精确度。

在 snownlp 中，通过 train() 和 save() 两个函数把模型训练和保存之后，就能实现扩展默认字典的功能了。此外，我在工作中还会利用这种方式增加 emoji 表情对应的情感倾向分析功能，以此来进一步提升 snownlp 分析情感倾向的准确度。

我将训练模型和保存训练后的模型的函数也写在这里供你参考，希望通过训练自己的模型，能够让你的产品分析更加准确。

复制代码

```
sentiment.train(neg2.txt,pos2.txt);  #    训练用户自定义正负情感数据集
sentiment.save('sentiment2.marshal');  # 保存训练模型
```

今天用到的代码，我都放在了 GitHub 上，你可以点击这个链接查看。

思考题

我给你留一道思考题，我在最后一段代码分别统计了正向和负向评价的数量，你能否根据这段代码统计一段文字中包含了多少个动词、多少个名词和多少个形容词呢？欢迎你在课程评论区留言，和我一起讨论。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 05 | 图像处理库：如何实现长图拼接？

下一篇 07 | 快速读写文件：如何实现跨文件的字数统计？

精选留言 (2)

写留言

Soul of the Drago...
2021-02-24

有个问题请教一下老师，我在思考题中用代码统计各种词性的数量，但每次统计的结果都不对，和实际数量相去甚远，不知道是什么原因。

展开

作者回复: 你好, 方便描述一下具体统计的是哪种词性吗? 导致这种结果的可能有连个原因, 一个是自动分词分的不够准确,导致了识别出现误差,还有一种原因是统计词性的英文缩写指定的不正确,例如名词就包含了很多详细分类,比如
名词分为以下子类:
n 名词

nr 人名
nr1 汉语姓氏
nr2 汉语名字
nrj 日语人名

可以搜索 jieba词性表 关键字对照分词的结果进行查看

1

陈东
2021-03-19

老师能分享你词性统计的思路吗?

展开

作者回复: 统计词性的一般思路是:
获取数据、数据预处理(去掉各种符号、空格)、词性标注(包含了jieba分词)、词频统计、数据再次处理(合并，计数，对统计结果格式调整)、数据保存到文件(一般为Excel)

按照这一顺序来实现词性统计

1